Additive regression modelling for zero-inflated Poisson distributions

Un modello additivo di regressione per la distribuzione zero-inflated Poisson

Monica Chiogna
Dipartimento di Scienze Statistiche
Università di Padova
monica@stat.unipd.it

Carlo Gaetan[1]
Dipartimento di Statistica
Università Ca' Foscari - Venezia
gaetan@unive.it

Abstract: In this paper, we develop nonparametric modelling of zero-inflated Poisson (ZIP) data. We model the nonparametric components of the regressors using penalized regression splines, so that inference can be accomplished by making use of the EM algorithm. The algorithm is illustrated by using an example from the literature.

Keywords: Penalized regression splines, EM algorithm.

1 Introduction and motivation

Depending on the context, count data may be modeled in a variety of ways, e.g. using Poisson, negative binomial, binomial, beta-binomial or hypergeometric distributions. However, real data often show an excess of some values compared their expected theoretical frequencies. In these cases, it is often said that the over-represented values are "inflated". A natural way to overcome this problem is to put a point mass ω at the inflated value. That is, with probability ω, we sample a degenerate distribution at the inflated value, and with probability $(1-\omega)$ we sample from a standard count distribution.

In this work, we treat a special inflation, i.e., the case of an excess of zeros in Poisson data. In particular, we develop nonparametric modelling of zero-inflated Poisson (ZIP) data. This relates to the work of Barry and Welsh (2002), who nonparametrically model ZIP data by following a two step process. First, they model whether the counts are zero or not. Then, they model the observed counts, conditionally on them being greater than 0. In both steps, they used a Generalized Additive Model (GAM). In our approach, the two steps process is overtaken by contemporary modelling the two components of the mixture via nonparametric or semiparametric functions. Our proposal is to construct the nonparametric components of the regressor using penalized regression splines, so that inference can be accomplished by making use of the EM algorithm.

[1] This work was supported by Ministero dell'Istruzione, dell'Università e della Ricerca Scientifica under grant n. 2004137478: *Trattamento statistico della complessità e dell'incertezza negli studi ambientali.*

2 The model

The ZIP distribution is defined as

$$p(y;\omega,\lambda) = \begin{cases} \omega + (1-\omega)\exp(-\lambda) & y = 0 \\ (1-\omega)\exp(-\lambda)\lambda^y/y! & y = 1, 2, \ldots \end{cases} \tag{1}$$

where $0 \leq \omega \leq 1$ and $\lambda > 0$.

In the ZIP regression models, $\omega = \omega_i = \omega(x_i)$ and $\lambda = \lambda_i = \lambda(x_i)$, where x_i is a vector of $p+q$ covariates, $i = 1, \ldots, n$. Lambert (1992) proposes the use the logit and log-linear link to model ω_i and λ_i, respectively, i.e.,

$$\log\left(\frac{\omega_i}{1-\omega_i}\right) = \sum_{j=1}^{q} \beta_j x_{ij}, \quad \log\lambda_i = \sum_{j=q+1}^{p+q} \beta_j x_{ij}, \quad i = 1, \ldots, n \tag{2}$$

where $\beta_j, j = 1, \ldots, p+q$, are unknown parameters to be estimated. In this formulation, the first q covariates of x_i are not necessarily different from the last p covariates.

Recently, Barry and Welsh (2002) have proposed to use GAMs for zero-inflated count data. Firstly, they model the presence of absence of counts via a GAM, i.e, they model $Y_i^* = \chi_{y>0}(Y_i)$, where χ_A is the indicator function of the set A, with the logistic link

$$\log\left(\frac{\omega_i}{1-\omega_i}\right) = \sum_{j=1}^{q} h_j(x_{ij}), \quad i = 1, \ldots, n.$$

Then, they model Y_i, conditional on $Y_i > 0$ by means of the truncated Poisson distribution:

$$p(y|y>0;\lambda_i) = \frac{\lambda_i^y \exp(-\lambda_i)}{y!\{1-\exp(-\lambda_i)\}},$$

with $\log\lambda_i = \sum_{j=q+1}^{p+q} h_j(x_{ij})$. The functions $h_j(\cdot)$, $j = 1, \ldots, p+q$, are smooth functions, estimated with linear smoothers by means of the backfitting algorithm (Hastie and Tibshirani, 1990).

While in certain applications this model is appropriate, the conditional approach introduces some bias if the model is not correctly specified, because the effects of the covariates on the probability of presence or absence might be not linear on the logistic scale. In our approach, we simultaneously model the two components of the ZIP distribution by extending model (2):

$$\log\left(\frac{\omega_i}{1-\omega_i}\right) = \sum_{j=1}^{q} h_j(x_{ij}), \quad \log\lambda_i = \sum_{j=q+1}^{p+q} h_j(x_{ij}), \quad i = 1, \ldots, n.$$

In order to estimate the nonparametric components of the regressors, we resort on penalized regression splines. In this setup, it is possible to show (see Wood (2000) and references therein) that the functions are estimated by maximizing, with respect to the parameter vector θ, the following penalized log-likelihood:

$$pl(\theta) = l(\theta) - \frac{1}{2}\sum_{i=1}^{p+q} \alpha_i \theta' S_i \theta,$$

where $l(\cdot)$ is the log-likelihood function, S_i are non-negative definite coefficient matrices defining the penalties and $\alpha_i > 0$ are the associated smoothing parameters, subject to the linear constraints $C\theta = 0$. Inference can be accomplished by adapting the EM algorithm. In fact, we can see (1) as a special case of the 2-component finite mixture model:

$$p(y; \omega, \lambda) = (1 - \omega)p_0(y; \lambda) + \omega p_1(y),$$

where $p_0(y; \lambda) = \exp(-\lambda)\lambda^y/y!$ and $p_1(y) = \chi_{\{0\}}(y)$. Now, let Z_i $i = 1, \ldots, n$, be n unobserved Bernoulli variables, such that $Pr(Z_i = 1) = \omega_i$, indicating which component the unit sample (y_i, x_i, z_i) comes from. The 'complete-data' log-likelihood is given by

$$l^C(\theta) = \sum_{i=1}^{n} \left\{ z_i \log\left(\frac{\omega_i}{1 - \omega_i}\right) + \log(1 - \omega_i) + (1 - z_i) \log p_0(y_i; \lambda_i) \right\}.$$

where, for the sake of simplicity, we have dropped the dependence of θ from ω_i and λ_i.

Our EM algorithm maximises the penalized log-likelihood by iterations of two steps: E-step and M-step. At stage k, given a current estimate θ_{k-1}, the E-step computes $S(\theta, \theta_{k-1})$, the expectation of $l^C(\theta)$ with respect the conditional distribution $p(z_1, \ldots, z_n | y_1, \ldots, y_n; \theta_k)$. The E-step is simple because the random variable Z_i given y_i is a Bernoulli random variable with probability of success

$$w(y_i; \theta) = \frac{\omega_i p_1(y_i)}{(1 - \omega_i)p_0(y_i; \lambda_i) + \omega_i p_1(y_i)}$$

and we have

$$\begin{aligned} S(\theta, \theta_{k-1}) &= \sum_{i=1}^{n} \left\{ w(y_i; \theta_{k-1}) \log\left(\frac{\omega_i}{1 - \omega_i}\right) \right. \\ &+ \left. \log(1 - \omega_i) + (1 - w(y_i; \theta_{k-1})) \log p_0(y_i; \lambda_i) \right\}. \end{aligned} \tag{3}$$

The M-step chooses $\theta = \theta_k$ to maximise

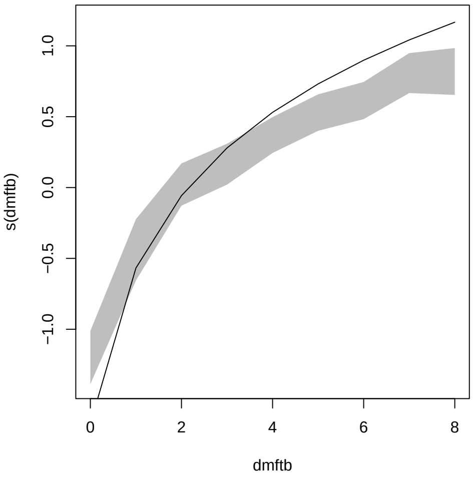
$$S(\theta, \theta_{k-1}) - \frac{1}{2} \sum_{i=1}^{p+q} \alpha_i \theta' S_i \theta,$$

subject to the linear constraints $C\theta = 0$. A simple application of the Jensen inequality shows that the sequence $pl(\theta_k)$ is not decreasing. In the M-step, it is easy to see that the first quantity in the right hand-side of (3) is equivalent to a weighted log-likelihood of a binomial regression model. On the other side, the second term is equivalent to a weighted log-likelihood of a Poisson regression model. This reduces the computational effort required by the M-step of the algorithm quite substantially, and allows us to use an Iteratively Reweighted Least Squares (IRLS) algorithm to perform the M-step.

3 An example

For illustrative purposes, we re-analyse data from a prospective study of school children from an urban area of Belo Horizonte (Brazil), the Belo Horizonte caries prevention (BELCAP) study. These data, which are available at

Figure 1: *Estimated nonparametric component for* `dmftb` *with approximated* 95% *confidence band. The solid line represents the original transformation.*



`http://www.blackwellpublishers.co.uk/rss`

have been previously analyzed by Böhning *et al.* (1999), by making use of a fully parametric ZIP regression model. We refer to the authors for a detailed description of the study and of the available data, In particular, one of the covariates (named by the authors DMFT1 and, in the following, `dmftb`) was log transformed. Here, we use the more flexible semiparametric model simply to validate the authors' choice on the type of transformation employed. We therefore fit to the data the semiparametric analogue of the ZIP model reported in Table 1 of the above mentioned paper. Figure 1 shows the estimated nonparametric component. The solid line indicates the transformation adopted by the authors, i.e. $\log(\texttt{dmftb} + 0.5)$. This seems to quite agree with indications deriving from the nonparametric analysis.

References

Barry S.C. and Welsh A. (2002) Generalized additive modelling and zero inflated count data, *Ecological Modelling*, 157, 179–188.

Böhning D., Dietz E., Schlattmann P., Mendonça L. and Kirchner U. (1999) The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology, *Journal of the Royal Statistical Society, Series A*, 162, 195–209.

Hastie T. and Tibshirani R. (1990) *Generalized Additive Models*, Chapman & Hall, London.

Lambert D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, 34, 1–14.

Wood S.N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties, *Journal of the Royal Statistical Society, Series B*, 62, 413–428.